

55. (Amended) A test kit for determining if BS322 antigen or anti-BS322 antibody is present in a test sample, said kit comprising:

sub  
FO  
a container containing at least one BS322 polypeptide having at least 95% identity over the entire length of a sequence selected from the group consisting of SEQUENCE ID NO: 25, SEQUENCE ID NO: 26, SEQUENCE ID NO: 27, and SEQUENCE ID NO: 28.

### REMARKS

The Examiner has maintained the rejection under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph. Applicant respectfully disagrees.

BS322 is a tissue specific putative transcription factor in breast tissue found by SEREX (Serological analysis of recombinant tumor cDNA expression libraries). Applicant submits Exhibit A, ["Homo sapiens similar to breast cancer antigen NY-BR-1 (LOC91074), mRNA"] and Exhibit B, [Jäger, D., *et al.*, "Identification of a Tissue-specific Putative Transcription Factor in Breast Tissue by Serological Screening of a Breast Cancer Library", *Cancer Research*, 61:2055-2061 (2001)], to illustrate this point. Exhibit A shows the homology between BS322 and the molecule NY-BR-1. As evidenced from this Exhibit, BS322 and NY-BR-1 are the same molecule. Exhibit B shows that NY-BR-1 (BS322) is found in breast cancer tissues. Specifically, NY-BR-1 (BS322) is found in 21 of 25 breast cancers but only in 2 of 82 non-mammary tumors. Thus, BS322 clearly has utility as a diagnostic tool for the detection of breast cancer.

The Examiner further rejects claims 52-61 under 35 U.S.C. § 112, first paragraph, due to the "90% identity" language in the claims. Applicant has raised the percent identity to 95%.

Applicant further submits the software manual to the Wisconsin Sequence Analysis program, Version 8, publicly available from Genetics Computer Group, Madison, WI, as Exhibit C. Support for this submission is found on page 16, beginning

line 7. The manual provides the algorithm, parameters, parameter values and other information necessary to, accurately and consistently, calculate the percent identity. This manual indicates on pages 5-21, *inter alia*, that the software used the local homology algorithm of Smith and Waterman (Advances in Applied Mathematics 2; 482-489 (1981)).

The Examiner further rejects claims 62-69, 72-76 and 80 under 35 U.S.C. § 112 paragraph 2 due to the "epitope" language percent in the claims.

The methods for identifying epitopes in a novel peptide sequence are well known and described in both the scientific, commercial, and patent literature. For example, M. H. Van Regenmortel describes how to predict epitopes from the primary sequence of a protein. (See "Protein structure and antigenicity", *Int J Rad Appl Instrum B.*, 14(4)277-80, 1987.)

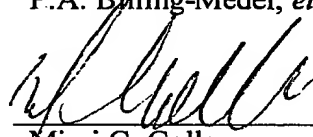
Further, Perkin-Elmer Biosystems, a major provider of DNA sequencing and peptide synthesizing instruments has established a public website which describes how to select peptides which reflect the epitopes of a protein. (See [http://www.pebio.com/pa/340913/html/chapt2.html#Choosing the Epitope.](http://www.pebio.com/pa/340913/html/chapt2.html#Choosing%20the%20Epitope)) This electronic publication was posted in 1996 and basically describes the process employed by the inventors of the current patent application.

In addition, patent application PCT/US97/00485 describes in detail how to identify epitopes from peptide sequences. The sequence can be scanned for hydrophobicity and hydrophilicity values by the method of Hopp, *Prog. Clin. Biol. Res.* 172B: 367-377 (1985) or the method of Cease et al, *J. Exp. Med.* 164: 1779-1784 (1986) or the method of Spouge et al, *J. Immunol.* 138: 204-212 (1987). Commercial software programs to implement these methods are available.

CONCLUSION

In view of the aforementioned remarks, Applicant respectfully submits that the above-referenced application is now in a condition for allowance and Applicant respectfully requests that the Examiner withdraw all outstanding objections and rejections and passes the application to allowance.

Respectfully submitted,  
P.A. Billing-Medel, *et al.*



---

Mimi C. Goller  
Registration No. 39,046  
Attorney for Applicants

ABBOTT LABORATORIES  
D-0377/AP6D-2  
100 Abbott Park Road  
Abbott Park, Illinois 60064-6050  
Phone: (847) 935-7550  
Fax: (847) 938-2623

**Version with Markings to Show Changes Made**

52. (Amended) A [BS322] purified polypeptide, having at least [90%] 95% identity over the entire length of a sequence selected from the group consisting of [SEQ ID NOS: 25-28] SEQUENCE ID NO: 25, SEQUENCE ID NO: 26, SEQUENCE ID NO: 27, and SEQUENCE ID NO: 28.

55. (Amended) A test kit for determining if BS322 antigen or anti-BS322 antibody is present in a test sample, said kit comprising:  
a container containing at least one BS322 polypeptide having at least [90%] 95% identity over the entire length of a sequence selected from the group consisting of [SEQ ID NOS: 24-28] SEQUENCE ID NO: 25, SEQUENCE ID NO: 26, SEQUENCE ID NO: 27, and SEQUENCE ID NO: 28.

#21  
HP  
3/18-02  
attachment

EXHIBIT A

RECEIVED

MAR 13 2002

TECH CENTER 1600/2900

Alignments

>gi|16156644|ref|XM 035844.3| Homo sapiens similar to breast cancer antigen  
NY-BR-1 (LOC91074),

mRNA

Length = 4408

Score = 2266 bits (1143), Expect = 0.0

Identities = 1143/1143 (100%)

Strand = Plus / Plus

Query: 1197 aggtttctcacactcatgaaaatgaaaattatctcttacatgaaaattgcatgttgaaaa 1256  
|||||

Sbjct: 3200 aggtttctcacactcatgaaaatgaaaattatctcttacatgaaaattgcatgttgaaaa 3259

Query: 1257 aggaaattgccatgctaaaactggaaatagccacactgaaacaccaataaccaggaaaagg 1316  
|||||

Sbjct: 3260 aggaaattgccatgctaaaactggaaatagccacactgaaacaccaataaccaggaaaagg 3319

Query: 1317 aaaataaatactttgaggacattaagatttttaaagaaaagaatgctgaacttcagatga 1376  
|||||

Sbjct: 3320 aaaataaatactttgaggacattaagatttttaaagaaaagaatgctgaacttcagatga 3379

Query: 1377 ccctaaaactgaaagaggaatcattaactaaaagggcatctcaatatagtgggcagctta 1436  
|||||

Sbjct: 3380 ccctaaaactgaaagaggaatcattaactaaaagggcatctcaatatagtgggcagctta 3439

Query: 1437 aagttctgatagctgagaacacaatgctcacttctaaattgaaggaaaaacaagacaaag 1496  
|||||

Sbjct: 3440 aagttctgatagctgagaacacaatgctcacttctaaattgaaggaaaaacaagacaaag 3499

Query: 1497 aaatactagaggcagaaattgaatcacaccatcctagactggcttctgctgtacaagacc 1556  
|||||

Sbjct: 3500 aaatactagaggcagaaattgaatcacaccatcctagactggcttctgctgtacaagacc 3559

Query: 1557 atgatcaaattgtgacatcaagaaaaagtcaagaacctgctttccacattgcaggagatg 1616  
|||||

Sbjct: 3560 atgatcaaattgtgacatcaagaaaaagtcaagaacctgctttccacattgcaggagatg 3619

Query: 1617 cttgtttgcaaagaaaaatgaatggtgatgtgagtagtacgatataataacaatgaggtgc 1676  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 3620 cttgtttgcaaagaaaaatgaatggtgatgtgagtagtacgatataataacaatgaggtgc 3679

Query: 1677 tccatcaaccacttttctgaagctcaaaggaaatccaaaagcctaaaaattaatctcaatt 1736  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 3680 tccatcaaccacttttctgaagctcaaaggaaatccaaaagcctaaaaattaatctcaatt 3739

Query: 1737 atgcaggagatgctctaagagaaaatacattggtttcagaacatgcacaaagagaccaac 1796  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 3740 atgcaggagatgctctaagagaaaatacattggtttcagaacatgcacaaagagaccaac 3799

Query: 1797 gtgaaacacagtgtcaaataaggaagctgaacacatgtatcaaaacgaacaagataatg 1856  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 3800 gtgaaacacagtgtcaaataaggaagctgaacacatgtatcaaaacgaacaagataatg 3859

Query: 1857 tgaacaaacacactgaacagcaggagtctctagatcagaaattatttcaactacaaagca 1916  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 3860 tgaacaaacacactgaacagcaggagtctctagatcagaaattatttcaactacaaagca 3919

Query: 1917 aaaatatgtggcttcaacagcaattagttcatgcacataagaaagctgacaacaaaagca 1976  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 3920 aaaatatgtggcttcaacagcaattagttcatgcacataagaaagctgacaacaaaagca 3979

Query: 1977 agataacaattgatattcattttcttgagaggaaaatgcaacatcatctcctaaaagaga 2036  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 3980 agataacaattgatattcattttcttgagaggaaaatgcaacatcatctcctaaaagaga 4039

Query: 2037 aaaatgaggagatatttaattacaataaccatttataaaaaccgtatatatcaatatgaaa 2096  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 4040 aaaatgaggagatatttaattacaataaccatttataaaaaccgtatatatcaatatgaaa 4099

Query: 2097 aagagaaagcagaaacagaaaactcatgagagacaagcagtaagaaacttcttttgaga 2156  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 4100 aagagaaagcagaaacagaaaactcatgagagacaagcagtaagaaacttcttttgaga 4159

Query: 2157 aacaacagaccagatctttactcacaactcatgctaggaggccagtcctagcatcacctt 2216  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 4160 aacaacagaccagatctttactcacaactcatgctaggaggccagtcctagcatcacctt 4219

Query: 2217 atgttgaaaatcttaccaatagtcctgtgtcaacagaatacttattttagaagaaaaattc 2276  
 ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  
 Sbjct: 4220 atgttgaaaatcttaccaatagtcctgtgtcaacagaatacttattttagaagaaaaattc 4279

Query: 2277 atgatttcttctgaagcctacagacataaaataacagtgtgaagaattacttgttcacg 2336  
|||||  
Sbjct: 4280 atgatttcttctgaagcctacagacataaaataacagtgtgaagaattacttgttcacg 4339

Query: 2337 aat 2339  
|||  
Sbjct: 4340 aat 4342

>gi|13469728|gb|AF269087.1|AF269087 Homo sapiens breast cancer antigen NY-BR-1  
mRNA, complete cds  
Length = 4458

Score = 2266 bits (1143), Expect = 0.0  
Identities = 1143/1143 (100%)  
Strand = Plus / Plus

Query: 1197 aggtttctcacactcatgaaaatgaaaattatctcttacatgaaaattgcatgttgaaaa 1256  
|||||  
Sbjct: 3197 aggtttctcacactcatgaaaatgaaaattatctcttacatgaaaattgcatgttgaaaa 3256

Query: 1257 aggaaattgccatgctaaaactggaaatagccacactgaaacaccaataaccaggaaaagg 1316  
|||||  
Sbjct: 3257 aggaaattgccatgctaaaactggaaatagccacactgaaacaccaataaccaggaaaagg 3316

Query: 1317 aaaataaatactttgaggacattaagattttaaaagaaaagaatgctgaacttcagatga 1376  
|||||  
Sbjct: 3317 aaaataaatactttgaggacattaagattttaaaagaaaagaatgctgaacttcagatga 3376

Query: 1377 ccctaaaactgaaagaggaatcattaactaaaaggcatctcaatatagtgggcagctta 1436  
|||||  
Sbjct: 3377 ccctaaaactgaaagaggaatcattaactaaaaggcatctcaatatagtgggcagctta 3436

Query: 1437 aagttctgatagctgagaacacaatgctcacttctaaattgaaggaaaaacaagacaaag 1496  
|||||  
Sbjct: 3437 aagttctgatagctgagaacacaatgctcacttctaaattgaaggaaaaacaagacaaag 3496

Query: 1497 aaatactagaggcagaaattgaatcacaccatcctagactggcttctgctgtacaagacc 1556  
|||||  
Sbjct: 3497 aaatactagaggcagaaattgaatcacaccatcctagactggcttctgctgtacaagacc 3556

Query: 1557 atgatcaaattgtgacatcaagaaaaagtcaagaacctgctttccacattgcaggagatg 1616  
|||||  
Sbjct: 3557 atgatcaaattgtgacatcaagaaaaagtcaagaacctgctttccacattgcaggagatg 3616

Query: 1617 cttgtttgcaaagaaaaatgaatgttgatgtgagtagtacgatataacaatgaggtgc 1676  
|||||

Sbjct: 3617 cttgtttgcaaagaaaaatgaatggtgatgtgagtagtacgatataataacaatgaggtgc 3676

Query: 1677 tccatcaaccactttctgaagctcaaaggaaatccaaaagcctaaaaattaatctcaatt 1736  
 |||

Sbjct: 3677 tccatcaaccactttctgaagctcaaaggaaatccaaaagcctaaaaattaatctcaatt 3736

Query: 1737 atgcaggagatgctctaagagaaaatacattggtttcagaacatgcacaaagagaccaac 1796  
 |||

Sbjct: 3737 atgcaggagatgctctaagagaaaatacattggtttcagaacatgcacaaagagaccaac 3796

Query: 1797 gtgaaacacagtgtcaaataaggaagctgaacacatgtatcaaaacgaacaagataatg 1856  
 |||

Sbjct: 3797 gtgaaacacagtgtcaaataaggaagctgaacacatgtatcaaaacgaacaagataatg 3856

Query: 1857 tgaacaaacacactgaacagcaggagtctctagatcagaaattatttcaactacaaagca 1916  
 |||

Sbjct: 3857 tgaacaaacacactgaacagcaggagtctctagatcagaaattatttcaactacaaagca 3916

Query: 1917 aaaatatgtggcttcaacagcaattagttcatgcacataagaaagctgacaacaaaagca 1976  
 |||

Sbjct: 3917 aaaatatgtggcttcaacagcaattagttcatgcacataagaaagctgacaacaaaagca 3976

Query: 1977 agataacaattgatattcattttcttgagaggaaaatgcaacatcatctcctaaaagaga 2036  
 |||

Sbjct: 3977 agataacaattgatattcattttcttgagaggaaaatgcaacatcatctcctaaaagaga 4036

Query: 2037 aaaatgaggagatatttaattacaataaccatttataaaacggtatatatcaatatgaaa 2096  
 |||

Sbjct: 4037 aaaatgaggagatatttaattacaataaccatttataaaacggtatatatcaatatgaaa 4096

Query: 2097 aagagaaagcagaaacagaaaactcatgagagacaagcagtaagaaacttcttttggaga 2156  
 |||

Sbjct: 4097 aagagaaagcagaaacagaaaactcatgagagacaagcagtaagaaacttcttttggaga 4156

Query: 2157 aacaacagaccagatctttactcacaactcatgctaggaggccagtcctagcatcacctt 2216  
 |||

Sbjct: 4157 aacaacagaccagatctttactcacaactcatgctaggaggccagtcctagcatcacctt 4216

Query: 2217 atgttgaaaatcttaccatagctctgtgtcaacagaatacttattttagaagaaaaattc 2276  
 |||

Sbjct: 4217 atgttgaaaatcttaccatagctctgtgtcaacagaatacttattttagaagaaaaattc 4276

Query: 2277 atgatttcttctgaagcctacagacataaaataacagtgtgaagaattacttggttcacg 2336  
 |||

Sbjct: 4277 atgatttcttctgaagcctacagacataaaataacagtgtgaagaattacttggttcacg 4336



Query: 2337 aat 2339

|||

Sbjct: 4337 aat 4339

# Identification of a Tissue-specific Putative Transcription Factor in Breast Tissue by Serological Screening of a Breast Cancer Library<sup>1</sup>

Dirk Jäger, Elisabeth St ckert, Ali O. Güre, Matthew J. Scanlan, Julia Karbach, Elke Jäger, Alexander Knuth, Lloyd J. Old, and Yao-Tseng Chen<sup>2</sup>

Weill Medical College of Cornell University, New York, New York 10021 [D. J., Y.-T. C.]; Medizinische Klinik, Hämatologie-Onkologie, Krankenhaus Nordwest, 60488 Frankfurt, Germany [J. K., E. J., A. K.]; and Ludwig Institute for Cancer Research, New York Branch, Memorial Sloan-Kettering Cancer Center, New York, New York 10021 [E. S., A. O. G., M. J. S., L. J. O., Y.-T. C.]

## ABSTRACT

Application of SEREX (serological analysis of recombinant tumor cDNA expression libraries) to different tumor types has led to the identification of several categories of human tumor antigens. In this study, the analysis of a breast cancer library with autologous patient serum led to the isolation of seven genes, designated *NY-BR-1* through *NY-BR-7*. *NY-BR-1*, representing 6 of 14 clones isolated, showed tissue-restricted mRNA expression in breast and testis but not in 13 other normal tissues tested. Among tumor specimens, *NY-BR-1* mRNA expression was found in 21 of 25 breast cancers but in only 2 of 82 nonmammary tumors. Structural analysis of *NY-BR-1* cDNA and the corresponding genomic sequences in the recently released working draft of human genome indicated that *NY-BR-1* is composed of 37 exons and has an open reading frame of 4.0–4.2 kb, encoding a peptide of *M<sub>r</sub>* 150,000–160,000. A bipartite nuclear localization signal motif indicates a nuclear site for *NY-BR-1*, and the presence of a bZIP site (DNA-binding site followed by leucine zipper motif) suggests that *NY-BR-1* is a transcription factor. Additional structural features include five tandem ankyrin repeats, implying a role for *NY-BR-1* in protein-protein interactions. *NY-BR-1* thus represents a breast tissue-specific putative transcription factor with autoimmunogenicity in breast cancer patients. In addition to *NY-BR-1*, a homologous gene, *NY-BR-1.1*, was identified in this study. *NY-BR-1.1* shares 54% amino acid homology with *NY-BR-1* and also shows tissue-restricted mRNA expression. However, unlike *NY-BR-1*, *NY-BR-1.1* mRNA is expressed in brain, in addition to breast and testis. The exon structure of *NY-BR-1.1* remains to be defined. Using human genome database, *NY-BR-1* was localized to chromosome 10p11–p12, and *NY-BR-1.1* was tentatively localized to chromosome 9.

## INTRODUCTION

Whether immunological factors play a role in the development, growth, and progression of human breast cancer remains a critical unresolved issue. The lymphocyte infiltrates frequently associated with breast cancer (1–5), particularly the intense T- and B-cell infiltrates in medullary carcinoma (6–8), and the reactive changes in the draining lymph nodes of breast cancer patients (9, 10) are consistent with the idea of immune recognition in breast cancer. However, efforts to relate the lymphocyte infiltrate and lymph node changes with prognosis have not yielded conclusive evidence for such an association (11). The search for breast cancer antigens that elicit humoral or cellular immune reactions in breast cancer patients also has a long history, from evidence for immune responses against the murine mammary tumor virus (12) and delayed hypersensitivity and humoral immunity against T/Tn antigens (13), to more recent findings of antibody and T-cell responses to p53 (14) and HER-2/neu (15, 16).

One major challenge confronting the analysis of autologous im-

mune responses in breast cancer, however, is the well-recognized difficulty of establishing breast cancer cell lines as targets for immunological analysis. This is in contrast to the relative ease of establishing lines from melanoma, renal cancer, and other tumor types. For this reason, the analysis of the human T-cell response against melanoma and the molecular identification of the antigens eliciting these responses are far more advanced in melanoma (17–19) than in breast cancer.

The recent development of SEREX,<sup>3</sup> a general method to analyze the humoral immune response of cancer patients that does not require autologous tumor cell lines, provides a powerful new way to dissect the immune response to breast cancer. Our initial application of SEREX to breast cancer led to the identification of p33ING1, encoded by a putative tumor suppressor gene in breast cancer, as an immunogenic breast cancer antigen. In addition, CT antigens, shown previously to be immunogenic antigens in other tumor types, were identified (20). In the present study, we have continued our effort to define breast cancer antigens by SEREX. Of the panel of antigens identified, a highly restricted breast autoimmunogenic differentiation antigen, *NY-BR-1*, was identified and characterized.

## MATERIALS AND METHODS

**Tumor Tissue and Cell Lines.** The BR17 tumor sample was derived from a s.c. metastasis of a 60-year-old female patient at Krankenhaus Nordwest. The patient had an unusually favorable history with metastatic ductal carcinoma of the breast. Breast cancer cell lines and cell lines of other tumor types were obtained from the repository maintained at the Ludwig Institute for Cancer Research, New York Branch at the Memorial Sloan-Kettering Cancer Center. Tumor tissues were obtained from the Departments of Pathology at The New York Presbyterian Hospital and the Memorial Sloan-Kettering Cancer Center.

**RNA Extraction and Construction of cDNA Expression Library.** Total RNA was extracted from the BR17 breast cancer sample by conventional CsCl-guanidine thiocyanate gradient method. A cDNA library was constructed in a λ-ZAP Express vector, using a commercial cDNA library kit (Stratagene).

**Immunoscreening of the cDNA Library.** The unamplified cDNA expression library was screened with the autologous serum at 1:200 dilution. The screening procedure was as described previously (21). Briefly, the serum was diluted 1:10, preabsorbed with phage-transfected *Escherichia coli* lysate, further diluted to 1:200, and incubated overnight at room temperature with the nitrocellulose membranes (Schleicher & Schuell) containing the phage plaques at a density of 4000–5000 pfu/130-mm plate. After washing, the filters were incubated with alkaline phosphatase-conjugated goat antihuman Fcγ secondary antibodies, and the reactive phage plaques were visualized by incubating with 5-bromo-4-chloro-3-indolyl-phosphate and nitroblue tetrazolium.

**Sequence Analysis of the Reactive Clones.** The reactive clones were subcloned, purified, and *in vivo* excised to pBK-CMV plasmid forms (Stratagene). Plasmid DNA was prepared by using the Wizard Miniprep DNA Purification System (Promega). The inserted DNA was evaluated by *EcoRI*-*XbaI* restriction mapping, and clones representing different cDNA inserts were sequenced. The sequencing reactions were performed by the DNA Sequencing

Received 8/25/00; accepted 12/28/00.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

<sup>1</sup> Supported by NIH Grant CA-68024 and by the Cancer Research Institute/Rhea (Rose Marie) Finell Memorial Fellowship for breast cancer research.

<sup>2</sup> To whom requests for reprints should be addressed, at Department of Pathology, Weill Medical College, Cornell University, 1300 York Avenue, New York, NY 10021. E-mail: ytchen@med.cornell.edu.

<sup>3</sup> The abbreviations used are: SEREX, serological analysis of recombinant tumor cDNA expression libraries; CT, cancer testis; EST, expressed sequence tag; RT-PCR, reverse transcription-PCR; RACE, rapid amplification of cDNA ends; ORF, open reading frame; CD, cluster of differentiation; pfu, plaque-forming unit.

Service at Cornell University (Ithaca, NY) using Applied Biosystems PRISM (Perkin-Elmer) automated sequencers. DNA and amino acid sequences were compared with sequences in the GenBank and the EST databases using the BLAST program. Genes identical to entries in the GenBank were classified as known genes, whereas those that shared sequence identity only to ESTs and those which have no identity in either GenBank or EST databases were designated as unknown genes.

**RT-PCR.** To evaluate the mRNA expression pattern of the cloned cDNA in normal and malignant tissues, total RNA was extracted from breast cancer cell lines and tumor specimens by the conventional CsCl-guanidine thiocyanate gradient method, and normal tissue RNA was obtained commercially (Clontech). Gene-specific oligonucleotide primers were designed to amplify cDNA segments of 300–600 bp in length, with the estimated primer melting temperature in the range of 65–70°C (see Figs. 2 and 4 for specific primer sequences). All primers were synthesized commercially (Operon Technologies, Alameda, CA). RT-PCR was performed using 30 amplification cycles in a thermal cycler (Perkin-Elmer) at an annealing temperature of 60°C, and the products were analyzed by 1.5% gel electrophoresis and ethidium bromide visualization.

**Rapid Amplification of cDNA Ends.** RACE reactions (5'-RACE and 3'-RACE) were performed using gene-specific and adaptor-specific primers in conjunction with Marathon-Ready normal testis cDNA and AmpliTaq Gold polymerase (Perkin-Elmer). Products were ligated into the PCR-direct cloning vector pGEMT plasmid and analyzed by restriction mapping and sequencing.

**Hybridization Screening of a Testicular Library.** A commercially obtained testis cDNA expression library (Stratagene) was screened using a NY-BR-1 PCR product as a probe (see Fig. 2 for primer sequences), as described in the Stratagene manual. Briefly, a total of  $5 \times 10^4$  pfu/150-mm plate were transferred to nitrocellulose membranes (Schleicher & Schuell), the membranes were submerged in denaturation solution (1.5 M NaCl and 0.5 M NaOH) for 5 min, transferred into neutralization solution (1.5 M NaCl and 0.5 M Tris-HCl) for 5 min, and then rinsed in 0.2 M Tris-HCl and  $2 \times$  SSC. The membranes were hybridized to a  $^{32}$ P-labeled DNA probe at high stringency condition (68°C, aqueous buffer) and washed at high stringency condition. Positive clones were subcloned, purified, and *in vivo* excised to pBK-CMV plasmid forms as described above.

## RESULTS

A total of  $7 \times 10^5$  pfu from the BR17 cDNA library were screened using autologous BR17 serum at 1:200 dilution. Fourteen reactive clones were purified and sequenced. Comparison to GenBank and EST database revealed that these 14 clones were derived from seven distinct genes, two known and five unknown. These genes, designated NY-BR-1 through NY-BR-7, are described in Table 1. Four clones were derived from the two known genes, *PBK-1* (BR17-76, BR17-118, and BR17-137) and *TI-227* (BR17-100). *PBK-1* and *TI-227* are universally expressed genes, because ESTs derived from these two genes have been reported in many different normal tissues. Of the

remaining clones, NY-BR-4 through NY-BR-7, represented by one clone each, were also universally expressed based on comparison to EST databank entries. The six remaining clones, BR17-1a, BR17-8, BR17-35b, BR17-44a, BR17-44b, and BR17-128, were derived from the same unknown gene, NY-BR-1. Three matching cDNA sequences for NY-BR-1 were found in the EST database, two derived from breast cancer (accession numbers AI951118 and AW373574), and the third (accession number AW170035) derived from a pooled tissue source (testis, fetal lung, and B cell), suggesting a possible tissue-restricted expression of NY-BR-1 mRNA (see below).

**Structural Analysis of NY-BR-1 cDNA.** Compilation of the six NY-BR-1 cDNA clones revealed a cDNA sequence of 1464 bp. Analysis showed a continuous ORF throughout this sequence, indicating that this is a partial cDNA sequence, truncated at both 5' and 3' ends. Comparison with the EST entry AW170035 (446 bp) revealed 100% sequence identity in the 89 bp overlapping the 5' sequence, with the EST entry extending 357 bp further in its 3' sequence than NY-BR-1 cDNA clones. Sequences of the other two EST entries (AI951118 and AW373574) are contained within NY-BR-1. Combining the EST sequence with the cloned NY-BR-1 sequence allowed the definition of the translational termination codon, with a 3' untranslated region of 333 bp.

To complete the missing 5' cDNA sequence, a testicular library was screened using a NY-BR-1 PCR product as a probe. One of the clones isolated during this screening extended the 5' sequence of NY-BR-1 1346 bp but did not provide a definite translation initiation site. On the basis of this cDNA sequence, a 5' RACE-PCR was performed, and the PCR product was cloned into the pGEMT plasmid vector and sequenced. This 5'-RACE sequence extended the cDNA sequence 1292 bp further 5', with the longest ORF starting at the ATG codon at position 100. No stop codon was found in the 99-bp 5' sequence, suggesting the possibility of additional 5' coding sequence in NY-BR-1. However, repeated 5'-RACE using different nested-primer pairs and adaptor-ligated cDNA derived from different NY-BR-1 mRNA-positive tissues (testis and breast, see below) failed to extend the 5' cDNA sequence further.

The available NY-BR-1 cDNA has a 4125-bp coding sequence and a 333-bp 3'-untranslated segment (submitted to GenBank, accession number AF269087). The predicted amino acid sequence from the possible ATG initiation codon (nucleotide position 100) is shown in Fig. 1. Motif analysis of the amino acid sequence using PROSITE and Pfam search programs identified a bipartite nuclear localization signal motif at amino acid position 17–34, suggesting that NY-BR-1 is a nuclear protein. Five tandem ankyrin repeats were also identified, located at amino acid positions 49–81, 82–114, 115–147, 148–180, and 181–213. The presence of a bZIP site (DNA-binding site followed by leucine zipper motif) at amino acid position 1077–1104 suggests that this nuclear protein functions as a transcription factor. Of interest, three additional repetitive elements were identified located between the ankyrin repeats and the NH<sub>2</sub>-terminal bZIP DNA-binding site (Fig. 1). The first repetitive element, consisting of 357 nucleotides (119 amino acids), is tandemly repeated three times, spanning amino acid residues 459–815. The second repetitive sequence, consisting of repeats of 11 amino acids, is located between amino acids 224 and 300 (seven repeats). The third repetitive sequence, consisting of only two repeats of 34 amino acids each, is located between amino acids 301–334 (Fig. 1).

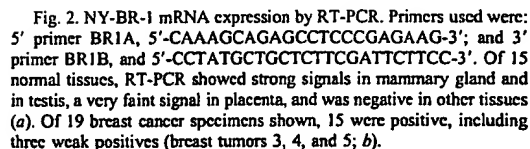
**mRNA Expression of NY-BR-1.** NY-BR-1 mRNA expression was tested in a panel of 15 different normal tissues (adrenal gland, fetal brain, lung, mammary gland, pancreas, placenta, prostate, thymus, uterus, ovary, brain, kidney, liver, colon, and testis). RT-PCR analysis showed a strong signal in mammary gland and testis and a very faint signal in placenta. All other tissues were negative (Fig. 2A).

Table 1 Clones identified by autologous SEREX screening of BR17 cDNA library

Designation	Clone	GenBank	Expression profile
NY-BR-1	BR17-128	Unknown	Expressed in normal breast and testis only (RT-PCR)
	BR17-1a		
	BR17-8		
	BR17-35b		
	BR17-44a		
	BR17-44b		
NY-BR-2	BR17-76	PBK-1	EST: ubiquitous
	BR17-118		
	BR17-137		
NY-BR-3	BR17-100	TI-227	EST: muscle, colon, endothelium, pancreas
NY-BR-4	BR17-91b	Unknown	EST: retina, cortex, fetal liver, spleen
NY-BR-5	BR17-115	Unknown	EST: uterus, melanocytes, fetal heart
NY-BR-6	BR17-117	Unknown	EST: tonsils, uterus, melanocytes, fetal heart
NY-BR-7	BR17-144	Unknown	EST: uterus, melanocytes, fetal heart

The expression of NY-BR-1 in tissue culture lines was also exam-

The available genomic sequence (GenBank AC067744) also allowed us to extend the 5' sequence of this gene beyond the cloned NY-BR-1 cDNA. Translation of the 5' genomic sequence using the previously assigned NY-BR-1 ORF led to the identification of a new translation initiation site 168 bp upstream to the previously predicted ATG initiation codon in NY-BR-1 cDNA (see text above and Fig. 1).



NY-BR-1	MTKRKKITNLNIQCAKRTALHMACVNGHEEVVTFVLDKRCQDVLGDSHRTPLMKALQCHQEACANILDSGADINLVVDVYGNALHYAVYSEILSVVA	100
NY-BR-1.1	M.....	3
NY-BR-1	KLLSHGAVIEVHHKASLTPLLSITKRSEIUEVFLIKNANANAVNKYKCTALMLAVCHSGSEIVGMLLQONVDVFAADICGVTAERYAVTCGFHIEHQ	200
NY-BR-1.1	T...Y.....Q.....A.Q.....K.T.....T.....F.ES.....I.E.....E..H.I...R...AAR.VNY..Q.	103
NY-BR-1	IMEYIRKLSKHNQNTNPEGTSAQTPDEAAPLAERTPDTEASLVEKTPDEAAPLVEKTPDEASLVEKTPDEAASLVEGTSQDKIQCLEKATSGKFEQSAAE	300
NY-BR-1.1	LL..H...P..P.....T.....R.....A.....G.....T.....	181
NY-BR-1	TPREITSAPAKETSEKFTWPAKGRPRKIAWEKKEDTPR-----EIMSPAK-ETSEKFTWAAKGRPRKIAWEKKETPVKTCGVARV-----	378
NY-BR-1.1	...K.LR.T.....S.....E.S...T..E..TSVKTECVAGVT.N.T.VL..G.SNMIAC.T.ETST.AS.N.DVSS.EPIFSLFGTTRTIENSQCTRV	281
NY-BR-1	---TSNKTIVLEKGRSKMIAC-----PTKESSTKASANDQRPSPSEKQEEDEYSCDSRSLFESSAKIQVCIPESIYQKVMENREVEEPPKPSAPKP	469
NY-BR-1.1	EEDFNLA..IIS.SAAQNYT.LPDATYQDKDIKTINHKE..M.....R.....W..G.....T.....M.....L.E.....	381
NY-BR-1	AIENQNSVPNKAFELKNEQTLRADPMFPESKQKDYENSWDSSELSCTVSQKDVCLPKATHQKEIDKINGKLEESPNKDGLLKATCGMKVSIPTKALEL	569
NY-BR-1.1	.V.....AQ...S.....D.....Y.....F.TLS.....V.....P...R...L.N.....	481
NY-BR-1	KDMQTFKAEPPGKPSAFEPATEMOKSVPNKALELNEQTWRA-----DEIL-PSESKQDYENSWDTELSCTVSQKDVCLPKAAHQKEIDKINGKLEG	663
NY-BR-1.1	..RE.....S.D.DGLLK.TCGRV.L.....D.R.LK.ESPDN.GL.K.TCGRVSLPNKALELKDRA..FKAAQM-F.SESK..DDEENSWDF..	578
NY-BR-1	SPVKDGLLKANCGMKVSIPTKALELMDMTFFKAEPEKPSAFEPATEMOKSVPNKALELNEQTLRADEILPSESKQDYENSWDSSELSCTVSQKDV	763
NY-BR-1.1	-SPLT..QNDVCLPKATHQ.EPDTLS-QKLE-.S.D.DGLLK.TCG.KI.L.....DER.FK.EDVSSV..TFLFGKPT--T.NSQS.KVEE..	673
NY-BR-1	LPKATHQKEIDKINGKLEESPDNDGFLKAPCRMKVSIPTKALELMDMTFFKAEPEKPSAFEPATEMOKSVPNKALELNEQTLRADPMFPESKQKQKVE	863
NY-BR-1.1	...T.KEGATKTVT...QOER.IGIIEE..QDQTNH..SE.G-----RK.DTKS-TSDSEIISVSDTQNY.CLP.A.Y---K-EIKTING.I.	754
NY-BR-1	ENSWDSSELSRETYSQKDVCPKATHQKEMDKISGKLEDSTSLSKILDVHSCERARELQKDHCEQRTGKMEQMKKFCVLKGLKSEAKEISOLENQVK	963
NY-BR-1.1	.SP-EKP.HF.PATEMOKS..NKGLEW-N.OTLR-A..T.....ALP...G..K.N...I.A.....N.....Q.E.....A.....	851
NY-BR-1	WEQELCSVRLTLNQEEKRNADILNEKINEELGRIEEQHRKELEVKQEQALRIQDIELKSVEENLQVSHYTHENENYLLHNCMLKKEIAHLKLEIA	1063
NY-BR-1.1	...V...K...P.....L..K...H...T.....T.....S..D.F.....V.....	946
NY-BR-1	TLKHQYQEKENKYPEDIKILKEKNAELQMTLKLKEESLTKRASQYSGQLVLAENTMLTSLKLEKQDKLEIAEIESHHPLASAVQDHDQIVTSRKSQ	1163
NY-BR-1.1	...H.V.....O.....L...QKTV.....RE.....	1011
NY-BR-1	EPAPHIAGDAQLQRIMNVDSSTIYNNEVLHQPLSEAQRKSKSLKINLYAGDALRENTLVSEHAQRDQRTQCQMKAEHMYQNEQDNNVNHKTEQESL	1263
NY-BR-1	DQKLFQLQSKNMWQQQLVHAHKKADNKSKITIDCHFLERKMQHLLKKEKNEIFNYNHLKNRIYOEKEKAETENS.	1341

Fig. 3. Comparison of the predicted amino acid sequences of NY-BR-1 and NY-BR-1.1. Identical sequences are shown as dots (•), and gaps are shown as dashes (—).

If this newly identified ATG is the true initiation site used *in vivo*, the NY-BR-1 polypeptide would contain 1397 amino acids, 56 residues longer than is depicted in Fig. 1 (additional NH<sub>2</sub> terminus sequence: MEEISAAAVKVVPGPERPSPFSQLVYTSNDSYTVHSGDLRKH-KAASRGQVRKLEK).

**Identification of NY-BR-1 Splice Variants.** Sequence comparison of the six SEREX-defined NY-BR-1 clones revealed that they were derived from two different splice variants. One variant contains an additional coding sequence of 111 bp (nucleotide nos. 3015–3125 of cloned NY-BR-1 cDNA, encoding amino acids 973–1009; see Fig. 1), which is absent in the other variant. Comparison with the genomic sequence confirmed that this results from an alternate splicing event, with the longer variant incorporating part of the intron 33 into exon 34 (i.e., exon 17 of the basic exon-intron framework described above). Key structural elements predicted above in the NY-BR-1 sequence are present in both splice variants, suggesting no apparent difference in biological function or subcellular localization.

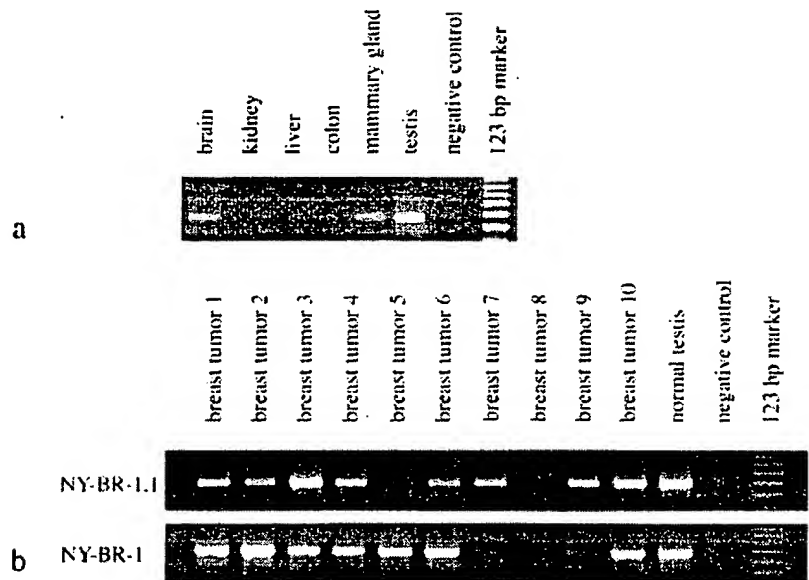
The expression of these two splice variants was evaluated using primers specific to the larger variant, as well as primers spanning the alternatively spliced exon. In the normal tissues analyzed, both variants showed strong expression in testis and breast by RT-PCR (but not in other tissues), with the larger variant being the dominant form in testis and the shorter variant dominant in breast. A selective group of 10 breast cancer samples were typed for these two splice forms, and results showed cotyping of the two variants (7 strong positive, 2 weaker positive, and 1 negative), with the shorter variant consistently being the predominant form.

**Isolation of a NY-BR-1 Homologue Gene.** Screening testicular cDNA library with a NY-BR-1 probe identified a cDNA encoding a new gene with homology to the NY-BR-1 gene. This clone, 3673 bp excluding the poly(A) tail, corresponded to nucleotides 1–3481 of the NY-BR-1 and showed 62% homology. A DNA database search revealed no sequence identity to GenBank "nr" database, and the new gene has been designated NY-BR-1.1 (submitted to GenBank, accession number AF269088). ORF analysis showed an ORF from nucle-

otide 641 to the end of the cloned sequence, with 54% homology to the putative NY-BR-1 protein sequence. The ATG initiation codon of NY-BR-1.1 is preceded by a 640-bp 5'-untranslated region with scattered stop codons. Comparison of the available NY-BR-1 and NY-BR-1.1 amino acid sequences is shown in Fig. 3. RT-PCR analysis for NY-BR-1.1 showed a tissue-restricted mRNA expression pattern distinct from NY-BR-1. Among six normal tissues examined, NY-BR-1.1 showed strong RT-PCR signal in testis, moderate signals in brain and breast tissues, and was negative in kidney, liver, and colon (Fig. 4A). Upon multiple repeated experiments, normal breast tissues showed either no or weak signals, consistently weaker than those observed in testis and often in brain, indicating a lower level of expression. NY-BR-1.1 expression was also examined in six breast cancer cell lines and 10 breast cancer specimens. One of six breast cancer cell lines was positive, in contrast to four of six for NY-BR-1. Eight of 10 breast cancer specimens were positive. In comparison with NY-BR-1 expression, 6 were positive for both, 1 was positive for NY-BR-1 only, 2 were positive for NY-BR-1.1 only, and 1 was negative for both (Fig. 4B). The strong expression in brain and low-level expression in normal breast and the lack of correlation in NY-BR-1 and NY-BR-1.1 expression in breast cancer lines and tissues indicate that these two gene products have a clearly distinct expression pattern.

**Genomic Sequence of NY-BR-1.1.** Comparison of the NY-BR-1.1 sequence with the released working draft of the human genome revealed one clone with sequence identity (GenBank AL359312). This clone was presumably derived from chromosome 9, indicating that NY-BR-1 and NY-BR-1.1 reside on two different chromosomes. The AL359312 genomic sequence does not contain the entire NY-BR-1.1 cDNA sequence, precluding the definition of NY-BR-1.1 exon-intron structure. However, at least three exons can be defined, which are the counterparts of the basic framework of exons 16, 17, and 18 in NY-BR-1. The exon-intron junctions of NY-BR-1 and NY-BR-1.1 are conserved in these exons.

Fig. 4. NY-BR-1.1 mRNA expression by RT-PCR. Primers used were: 5' primer BR-1.1A, 5'-TCTCATAGATGCTGGTGTGCTGATC-3'; and 3' primer BR-1.1B, and 5'-CCCAGACATTGAATTTTGGCAGAC-3'. Of six normal tissues tested, RT-PCR showed a strong signal in testis, moderate signals in brain and mammary gland, and negative in kidney, liver, and colon (a). Of 10 breast cancer specimens, 8 were positive for NY-BR-1.1 (b). Comparing the NY-BR-1 and NY-BR-1.1 expression, 7 of 10 cotyped (6 positive and 1 negative), whereas two were positive for NY-BR-1 only (tumors 7 and 9), and one was positive for NY-BR-1 only (tumor 5).



## DISCUSSION

The SEREX approach has proved to be a very powerful tool to identify tumor antigens (20–30). SEREX-defined antigens have been classified into several categories, including differentiation antigens, CT antigens, mutational antigens, amplified/overexpressed antigens, splice variant antigens, and retroviral antigens (31). The expression of NY-BR-1 in breast tissue and testis but not in other normal tissues indicates that NY-BR-1 belongs to the category of differentiation antigens.

Differentiation antigens are antigens that show expression in specific cell lineage(s) or at specific stages of differentiation in a particular cell lineage(s) (32). This category of antigens has been best studied in cells of lymphoid and hematopoietic derivation, starting with the definition of mouse cell surface differentiation antigens of lymphocytes, such as TL (33, 34), Thy-1 (35), and Lyt-2 (36). Application of hybridoma technology to the analysis of human cells resulted in the identification of a broad range of differentiation antigens, and this has led to the classification system for CD antigens (37). Most of the initial CD antigens were restricted differentiation antigens expressed in lymphocytes and other hematopoietic cells, e.g., CD1 through CD8 primarily restricted to T cells (38). The expression of differentiation antigens in normal cells is generally preserved in their neoplastic counterparts, and this feature has made these antigens useful markers in the immunopathological differential diagnosis of cancer. The best example of this is again in the hematopoietic/lymphocytic lineages, in which the antigenic profile of the neoplastic cells provided the foundation for the classification of leukemia/lymphoma (39). In addition, these antigens can be targets for specific immunotherapy, and anti-CD20 antibody, recognizing a B-cell differentiation antigen, represents the first monoclonal antibody approved by the Food and Drug Administration for cancer immunotherapy (40).

In addition to cells of hematopoietic origin, the melanocyte, a specialized cell type in the neuroectodermal lineage, has been found to express a number of well-characterized differentiation antigens, most of them associated with the melanin-synthesis pathway. Studies using polyclonal and monoclonal antibodies initially defined tyrosinase (41), gp75 (42), and gp100 (43). Recent efforts to identify melanoma antigens recognized by CD8+ and CD4+ T cells also identified these antigens as T-cell targets and further expanded the list of melanocyte differentiation antigens (44–48). Melan-A/MART-1,

identified by transfection-based T-cell epitope cloning as a CD8+ T-cell target (48, 49), and Rab38 (50), identified by SEREX analysis of melanoma, are two further examples of melanocyte differentiation antigens identified through these efforts. Similar to their hematopoietic counterparts, the melanocyte differentiation antigens have also been found useful in the clinical arena. Antibodies against gp100 and Melan-A/MART-1 are widely used to distinguish metastatic melanomas from other metastatic malignancies (51), and melanoma vaccine trials targeting these antigens are being actively pursued (see Ref. 52 for an example).

With regard to common epithelial tissue, a wide range of gene products with differential expression have been defined, e.g., cytokeratins (53, 54), mucin-related antigens (55), and hormonal receptors (56). However, with rare exceptions, none of these are exclusively expressed in only a single epithelial cell type, such as breast epithelium. In this regard, NY-BR-1 is of considerable interest because of its highly restricted expression pattern in normal tissue, i.e., breast and testis. The production of antibody probes for NY-BR-1 is essential to confirm breast specificity at the protein and cell levels. With regard to cancer vaccine development, the restricted expression of NY-BR-1 in normal breast and breast cancer makes it a highly attractive vaccine target. However, the presence of a homologous gene, *NY-BR-1.1*, that is expressed in brain is cause for concern, and it will be necessary to show that T-cell reactivity to NY-BR-1 can be generated without cross-recognition of NY-BR-1.1.

The predicted protein sequence of NY-BR-1 contains a DNA-binding site and a leucine zipper motif (bZIP). The bZIP motif characterizes the superfamily of eucaryotic DNA-binding transcription factors that contain a basic region mediating sequence-specific DNA-binding, followed by a leucine zipper required for dimerization (57, 58). It is thus most likely that NY-BR-1 is a transcription factor. Five ankyrin repeats are also present in the NY-BR-1 protein. Ankyrin repeat proteins carry out a wide variety of biological activities and are found in both cytoplasm and nucleus. The repeat motif has been recognized in >400 proteins, including cyclin-dependent kinase inhibitors, transcriptional regulators, cytoskeletal organizers, developmental regulators, and toxins (59). Thus, the ankyrin repeat in itself is not predictive of a specific cellular function or subcellular localization; rather, ankyrin repeats are thought to mediate a wide range of protein-protein interactions (59). In comparison to other ankyrin



repeat-containing proteins, NY-BR-1 is unique because of the other repetitive elements in its predicted protein sequence. These additional repetitive elements are not found in other sequences in the public protein databases, and their functional significance remains to be determined.

By comparing the cDNA sequence with the recently released working draft of the human genome, we were able to derive the following important information about *NY-BR-1*: (a) confirming the cDNA sequences of *NY-BR-1* and *NY-BR-1.1*; (b) mapping to chromosome 10p11-12 and *NY-BR-1.1* tentatively to chromosome 9; (c) definition of the exon-intron structure of *NY-BR-1*, a complex gene with 37 exons, and correlate exon structure to repeating peptide units; and (d) completion of the NH<sub>2</sub> terminus amino acid sequence of *NY-BR-1* that had defied our multiple cloning efforts and RACE analysis. On the other hand, the cDNA sequences of *NY-BR-1* and *NY-BR-1.1* genes from this study will certainly help the annotation of corresponding genome sequences. The current study thus provides a clear example of valuable data that can be achieved by interaction between the human genome project and other scientific fields.

## REFERENCES

- Shimokawara, I., Imamura, M., Yamanaka, N., Ishii, Y., and Kikuchi, K. Identification of lymphocyte subpopulations in human breast cancer tissue and its significance: an immunoperoxidase study with antihuman T- and B-cell sera. *Cancer (Phila.)*, 49: 1456-1464, 1982.
- Giorno, R. Mononuclear cells in malignant and benign human breast tissue. *Arch. Pathol. Lab. Med.*, 107: 415-417, 1983.
- Bhan, A. K., and DesMarais, C. L. Immunohistologic characterization of major histocompatibility antigens and inflammatory cellular infiltrate in human breast cancer. *J. Natl. Cancer Inst.*, 71: 507-516, 1983.
- Hurlimann, J., and Saraga, P. Mononuclear cells infiltrating human mammary carcinomas: immunohistochemical analysis with monoclonal antibodies. *Int. J. Cancer*, 35: 753-762, 1985.
- Gottlinger, H. G., Rieber, P., Gokel, J. M., Lohe, K. J., and Riethmuller, G. Infiltrating mononuclear cells in human breast carcinoma: predominance of T4+ monocyte cells in the tumor stroma. *Int. J. Cancer*, 35: 199-205, 1985.
- Ben-Ezra, J., and Sheibani, K. Antigenic phenotype of the lymphocytic component of medullary carcinoma of the breast. *Cancer (Phila.)*, 59: 2037-2041, 1987.
- Gaffey, M. J., Friserson, H. F., Jr., Mills, S. E., Boyd, J. C., Zarbo, R. J., Simpson, J. F., Gross, L. K., and Weiss, L. M. Medullary carcinoma of the breast. *Mod. Pathol.*, 6: 721-728, 1993.
- Hsu, S. M., Raine, L., and Nayak, R. N. Medullary carcinoma of breast: an immunohistochemical study of its lymphoid stroma. *Cancer (Phila.)*, 48: 1368-1376, 1981.
- Hirsch, S., Black, M. M., and Kwon, C. S. Ultrastructural characteristics of sinus histiocytic reaction in lymph nodes draining various stages of breast cancer. *Cancer (Phila.)*, 38: 807-817, 1976.
- Black, M. M., Barclay, T. H., and Hankey, B. F. Prognosis in breast cancer utilizing histologic characteristics of the primary tumor. *Cancer (Phila.)*, 36: 2048-2055, 1975.
- Fisher, E. R., Costantino, J., Fisher, B., and Redmond, C. Pathologic findings from the National Surgical Adjuvant Breast Project (Protocol 4). *Cancer (Phila.)*, 71: 2141-2150, 1993.
- Black, M. M., Zachrau, R. E., Shore, B., Moore, D. H., and Leis, H. P., Jr. Prognostically favorable immunogens of human breast cancer tissue: antigenic similarity to murine mammary tumor virus. *Cancer (Phila.)*, 35: 121-128, 1975.
- Springer, G. F. Immunoreactive T and Tn epitopes in cancer diagnosis, prognosis, and immunotherapy. *J. Mol. Med.*, 75: 594-602, 1997.
- Gnjatic, S., Cai, Z., Viguier, M., Chouaib, S., Guillet, J. G., and Chopin, J. Accumulation of the p53 protein allows recognition by human CTL of a wild-type p53 epitope presented by breast carcinomas and melanomas. *J. Immunol.*, 160: 328-333, 1998.
- Disis, M. L., Calenoff, E., McLaughlin, G., Murphy, A. E., Chen, W., Groner, B., Jeschke, M., Lydon, N., McGlynn, E., Livingston, R. B., et al. Existent T-cell and antibody immunity to HER-2/neu protein in patients with breast cancer. *Cancer Res.*, 54: 16-20, 1994.
- Disis, M. L., and Cheever, M. A. HER-2/neu protein: a target for antigen-specific immunotherapy of human cancer. *Adv. Cancer Res.*, 71: 343-371, 1997.
- Oettgen, H. F., Livingston, P. O., and Old, L. J. Immunotherapy by active specific immunization. In: S. H. V. DeVita and S. A. Rosenberg (eds.), *Biologic Therapy of Cancer*, Vol. 1, pp. 682-701. Philadelphia: J. B. Lippincott Co., 1991.
- Boon, T., and van der Bruggen, P. Human tumor antigens recognized by T lymphocytes. *J. Exp. Med.*, 183: 725-729, 1996.
- Kawakami, Y., and Rosenberg, S. A. Human tumor antigens recognized by T-cells. *Immunol. Res.*, 16: 313-339, 1997.
- Jager, D., Stockert, E., Scanlan, M. J., Gure, A. O., Jager, E., Knuth, A., Old, L. J., and Chen, Y. T. Cancer-testis antigens and ING1 tumor suppressor gene product are breast cancer antigens: characterization of tissue-specific ING1 transcripts and a homologous gene. *Cancer Res.*, 59: 6197-6204, 1999.
- Chen, Y. T., Scanlan, M. J., Sahin, U., Tureci, O., Gure, A. O., Tsang, S., Williamson, B., Stockert, E., Pfreundschuh, M., and Old, L. J. A testicular antigen aberrantly expressed in human cancers detected by autologous antibody screening. *Proc. Natl. Acad. Sci. USA*, 94: 1914-1918, 1997.
- Jager, E., Chen, Y. T., Drijfhout, J. W., Karbach, J., Ringhoffer, M., Jager, D., Arand, M., Wada, H., Noguchi, Y., Stockert, E., Old, L. J., and Knuth, A. Simultaneous humoral and cellular immune response against cancer-testis antigen NY-ESO-1: definition of human histocompatibility leukocyte antigen (HLA)-A2-binding peptide epitopes. *J. Exp. Med.*, 187: 265-270, 1998.
- Sahin, U., Tureci, O., Schmitt, H., Cochlovius, B., Johannes, T., Schmits, R., Stenner, F., Luo, G., Schobert, I., and Pfreundschuh, M. Human neoplasms elicit multiple specific immune responses in the autologous host. *Proc. Natl. Acad. Sci. USA*, 92: 11810-11813, 1995.
- Tureci, O., Sahin, U., Zwick, C., Koslowski, M., Scitz, G., and Pfreundschuh, M. Identification of a meiosis-specific protein as a member of the class of cancer/testis antigens. *Proc. Natl. Acad. Sci. USA*, 95: 5211-5216, 1998.
- Gure, A. O., Tureci, O., Sahin, U., Tsang, S., Scanlan, M. J., Jager, E., Knuth, A., Pfreundschuh, M., Old, L. J., and Chen, Y. T. SSX: a multigene family with several members transcribed in normal testis and human cancer. *Int. J. Cancer*, 72: 965-971, 1997.
- Gure, A. O., Altorki, N. K., Stockert, E., Scanlan, M. J., Old, L. J., and Chen, Y. T. Human lung cancer antigens recognized by autologous antibodies: definition of a novel cDNA derived from the tumor suppressor gene locus on chromosome 3p21.3. *Cancer Res.*, 58: 1034-1041, 1998.
- Scanlan, M. J., Chen, Y. T., Williamson, B., Gure, A. O., Stockert, E., Gordan, J. D., Tureci, O., Sahin, U., Pfreundschuh, M., and Old, L. J. Characterization of human colon cancer antigens recognized by autologous antibodies. *Int. J. Cancer*, 76: 652-658, 1998.
- Chen, Y. T., Gure, A. O., Tsang, S., Stockert, E., Jager, E., Knuth, A., and Old, L. J. Identification of multiple cancer/testis antigens by allogeneic antibody screening of a melanoma cell line library. *Proc. Natl. Acad. Sci. USA*, 95: 6919-6923, 1998.
- Tureci, O., Sahin, U., and Pfreundschuh, M. Serological analysis of human tumor antigens: molecular definition and implications. *Mol. Med. Today*, 3: 342-349, 1997.
- Brass, N., Heckel, D., Sahin, U., Pfreundschuh, M., Sybrecht, G. W., and Meese, E. Translation initiation factor eIF-4γ is encoded by an amplified gene and induces an immune response in squamous cell lung carcinoma. *Hum. Mol. Genet.*, 6: 33-39, 1997.
- Old, L. J., and Chen, Y. T. New paths in human cancer serology. *J. Exp. Med.*, 187: 1163-1167, 1998.
- Retting, W. J., and Old, L. J. Immunogenetics of human cell surface differentiation. *Annu. Rev. Immunol.*, 7: 481-511, 1989.
- Schlesinger, M., and Hurvitz, D. Differentiation of the thymus-leukemia (TL) antigen in the thymus of mouse embryos. *Isr. J. Med. Sci.*, 4: 1210-1215, 1968.
- Old, L. J. Cancer immunology: the search for specificity—G. H. A. Clowes Memorial lecture. *Cancer Res.*, 41: 361-375, 1981.
- Shiku, H., Kislacow, P., Bean, M. A., Takahashi, T., Boyse, E. A., Oettgen, H. F., and Old, L. J. Expression of T-cell differentiation antigens on effector cells in cell-mediated cytotoxicity *in vitro*. *J. Exp. Med.*, 141: 227-241, 1975.
- Cantor, H., and Boyse, E. A. Functional subclasses of T-lymphocytes bearing different Ly antigens. *J. Exp. Med.*, 141: 1376-1389, 1975.
- Bernard, A., Boumsell, L., and Hill, C. Joint report of the First International Workshop on Human Leukocyte Differentiation Antigens by the investigators of the participating laboratories. In: A. Berenard, L. Boumsell, J. Dausset, C. Milstein, and S. F. Schlossman (eds.), *Leukocyte Typing: Human Leukocyte Differentiation Antigens Selected by Monoclonal Antibodies*, pp. 9-142. New York: Springer-Verlag, 1984.
- Knapp, W., Dorken, B., Gilks, W. R., Rieber, E. P., Schmidt, R. E., Stein, H., and von dem Borne, A. E. G. Kr. (eds.). *Leukocyte Typing IV*, pp. 229-386. New York: Oxford University Press, 1989.
- Nathwani, B. N., Brynes, R. K., Lincoln, T., Taylor, C. R., and Hanseman, M. L. Classification of the non-Hodgkin's lymphomas. In: D. M. Knowles (ed.), *Neoplastic Hematopathology*, pp. 555-602. Baltimore: Williams and Wilkins, 1992.
- Grillo-Lopez, A. J., White, C. A., Varns, C., Shen, D., Wei, A., McClure, A., and Dallaire, B. K. Overview of the clinical development of rituximab: first monoclonal antibody approved for the treatment of lymphoma. *Semin. Oncol.*, 26: 66-73, 1999.
- Jimenez, M., Tsukamoto, K., and Hearing, V. J. Tyrosinases from two different loci are expressed by normal and by transformed melanocytes. *J. Biol. Chem.*, 266: 1147-1156, 1991.
- Mattes, M. J., Thomson, T. M., Old, L. J., and Lloyd, K. O. A pigmentation-associated, differentiation antigen of human melanoma defined by a precipitating antibody in human serum. *Int. J. Cancer*, 32: 717-721, 1983.
- Gown, A. M., Vogel, A. M., Hoak, D., Gough, F., and McNutt, M. A. Monoclonal antibodies specific for melanocytic tumors distinguish subpopulations of melanocytes. *Am. J. Pathol.*, 123: 195-203, 1986.
- Brichard, V., Van Pel, A., Wolfel, T., Wolfel, C., De Plaen, E., Lethe, B., Coulic, P., and Boon, T. The tyrosinase gene codes for an antigen recognized by autologous cytotoxic T lymphocytes on HLA-A2 melanomas. *J. Exp. Med.*, 178: 489-495, 1993.
- Wang, R. F., Robbins, P. F., Kawakami, Y., Kang, X. Q., and Rosenberg, S. A. Identification of a gene encoding a melanoma tumor antigen recognized by HLA-A31-restricted tumor-infiltrating lymphocytes [published erratum appears in *J. Exp. Med.*, 181: 1261, 1995]. *J. Exp. Med.*, 181: 799-804, 1995.
- Wang, R. F., Appella, E., Kawakami, Y., Kang, X., and Rosenberg, S. A. Identification of TRP-2 as a human tumor antigen recognized by cytotoxic T lymphocytes. *J. Exp. Med.*, 184: 2207-2216, 1996.
- Coulic, P. G., Brichard, V., Van Pel, A., Wolfel, T., Schneider, J., Traversari, C., Mattei, S., De Plaen, E., Lurquin, C., Szikora, J. P., Renauld, J.-C., and Boon, T. A

- new gene coding for a differentiation antigen recognized by autologous cytolytic T lymphocytes on HLA-A2 melanomas. *J. Exp. Med.*, 180: 35-42, 1994.
48. Kawakami, Y., Eliyahu, S., Delgado, C. H., Robbins, P. F., Rivoltini, L., Topalian, S. L., Miki, T., and Rosenberg, S. A. Cloning of the gene coding for a shared human melanoma antigen recognized by autologous T cells infiltrating into tumor. *Proc. Natl. Acad. Sci. USA*, 91: 3515-3519, 1994.
49. Van den Eynde, B. J., and van der Bruggen, P. T cell defined tumor antigens. *Curr. Opin. Immunol.*, 9: 684-693, 1997.
50. Jäger, D., Stockert, E., Jäger, E., Gürc, A., Scanlan, M. J., Knuth, A., Old, L. J., and Chen, Y.-T. Serological cloning of a melanocyte rab guanosine 5'-triphosphate-binding protein and a chromosome condensation protein from a melanoma complementary DNA library. *Cancer Res.*, 60: 3584-3591, 2000.
51. Jungbluth, A. A., Busam, K. J., Gerald, W. L., Stockert, E., Coplan, K. A., Iversen, K., MacGregor, D. P., Old, L. J., and Chen, Y. T. A103: an antimelanoma monoclonal antibody for the detection of malignant melanoma in paraffin-embedded tissues. *Am. J. Surg. Pathol.*, 22: 595-602, 1998.
52. Rosenberg, S. A., Yang, J. C., Schwartzentruber, D. J., Hwu, P., Marincola, F. M., Topalian, S. L., Restifo, N. P., Dudley, M. E., Schwarz, S. L., Spicess, P. J., Wunderlich, J. R., Parkhurst, M. R., Kawakami, Y., Seipp, C. A., Einhorn, J. H., and White, D. E. Immunologic and therapeutic evaluation of a synthetic peptide vaccine for the treatment of patients with metastatic melanoma. *Nat. Med.*, 4: 321-327, 1998.
53. Fuchs, E., Tyner, A. L., Giudice, G. J., Marchuk, D., Raychaudhury, A., and Rosenberg, M. The human keratin genes and their differential expression. *Curr. Top. Dev. Biol.*, 22: 5-34, 1987.
54. Moll, R. Cytokeratins in the histological diagnosis of malignant tumors. *Int. J. Biol. Markers*, 9: 63-69, 1994.
55. Denton, G., Sckowski, M., Spencer, D. I., Hughes, O. D., Murray, A., Denley, H., Tandler, S. J., and Price, M. R. Production and characterization of a recombinant anti-MUC1 scFv reactive with human carcinomas. *Br. J. Cancer*, 76: 614-621, 1997.
56. Ferguson, A. T., Lapidus, R. G., and Davidson, N. E. The regulation of estrogen receptor expression and function in human breast cancer. *Cancer Treat. Res.*, 94: 255-278, 1998.
57. Hurst, H. C. Transcription factors I: bZIP proteins. *Protein Profile*, 2: 101-168, 1995.
58. Ellenberger, T., Fass, D., Amaud, M., and Harrison, S. C. Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. *Genes Dev.*, 8: 970-980, 1994.
59. Sedgwick, S. G., and Smerdon, S. J. The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem. Sci.*, 24: 311-316, 1999.



## Exhibit C

## FUNCTION

BestFit makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the *local homology* algorithm of Smith and Waterman.

## DESCRIPTION

BestFit inserts gaps to obtain the optimal alignment of the best region of similarity between two sequences, and then displays the alignment in a format similar to the output from Gap. The sequences can be of very different lengths and have only a small segment of similarity between them. You could take a short RNA sequence, for example, and run it against a whole mitochondrial genome.

## SEARCHING FOR SIMILARITY

BestFit is the most powerful method in the Wisconsin Sequence Analysis Package™ for identifying the best region of similarity between two sequences whose relationship is unknown.

## EXAMPLE

The sequence gamma.seq contains an Alu family sequence somewhere in the first 500 bases. alu.seq contains a generic human Alu family repeat. The two sequences are aligned and the best segment of similarity is found with BestFit.

```
% bestfit
```

```
BESTFIT of what sequence 1 ? gamma.seq
```

```
      Begin (* 1 *) ?  
      End   (* 11375 *) ? 500  
      Reverse (* No *) ?
```

```
to what sequence 2 (* gamma.seq *) ? alu.seq
```

```
      Begin (* 1 *) ?  
      End   (* 207 *) ?  
      Reverse (* No *) ?
```

```
What is the gap creation penalty (* 5.00 *) ?
```

```
What is the gap extension penalty (* 0.30 *) ?
```

```
What should I call the paired output display file (* gamma.pair *)
```

```
Aligning .....-..
```

```
      Gaps:      3  
      Quality: 129.3  
      Quality Ratio: 0.625  
      % Similarity: 84.466  
      Length:   209
```

## - OUTPUT

Here is the output file. Notice how BestFit finds and displays only the best segments of similarity:

BESTFIT of: gamma.seq check: 6474 from: 1 to: 500

Human fetal beta globins G and A gamma  
from Shen, Slightom and Smithies, Cell 26; 191-203.  
Analyzed by Smithies et al. Cell 26; 345-353.

to: alu.seq check: 4238 from: 1 to: 207

HSREP2 from the EMBL data library

Human Alu repetitive sequence located near the insulin gene  
Dhruva D.R., Shenk T., Subramanian K.N.; "Integration in vivo into  
Simian virus 40 DNA of a sequence that resembles a certain family of  
genomic interspersed repeated sequences"; Proc. Natl. Acad. Sci. USA  
77:4514-4518 (1980). . . .

Symbol comparison table: Gcgcordisk:[Gcgcore.Data.Rundata]Swgapdna.Cmp  
CompCheck: 5234

Gap Weight:	5.000	Average Match:	1.000
Length Weight:	0.300	Average Mismatch:	-0.900

Quality:	129.3	Length:	209
Ratio:	0.625	Gaps:	3
Percent Similarity:	84.466	Percent Identity:	84.466

gamma.seq x alu.seq June 20, 1994 15:15 ..

```

137 AGACCAACCTGGCCAACATGGTGAAATCCCATCTCTAC.AAAAATACAAA 185
||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
1 AGACCAGCCTGGCCAACATGGTGAAACTCCATCTCTACTGAAAATACAAA 50

186 AATTAGACAGGCATGATGGCAAGTGCCTGTAATCCCAGCTACTTGGGAGG 235
||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
51 AATTAGCCAGGCATGGTGATGCGTGCCTGGAATCCCAGCTACTTAGGAGG 100

236 CTGAGGAAGGAGAATTGCTTGAACCTGGAAGGCAGGAGTTGCAGTGAGCC 285
||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
101 CTGAGACAGAAGAATCCCTTAAACCAAG.AGGTGGAGGTTGCAGTGAGCC 149

286 GAGATCATACCACTGCACTCCAGCCTGGGTGACAGAACAGAETCTGTCT 335
||||| ||||| ||||| ||||| ||||| ||||| ||||| |||||
150 GAGATCGCACGGCTGCACTCCAGCCT.GGTGACAGAGCGAGACTCCATCT 198

336 CAAAAAAAAA 344
199 CAAAAAAAAA 207

```

## RELATED PROGRAMS

When you want an alignment that covers the whole length of both sequences, use Gap. When you are trying to find only the best segment of similarity between two sequences, use BestFit. PileUp creates a multiple sequence alignment of a group of related sequences, aligning the whole length of all sequences. DotPlot displays the entire surface of comparison for a comparison of two sequences. GapShow displays the pattern of differences between two aligned sequences. PlotSimilarity plots the average similarity of two or more aligned sequences at each position in the alignment. Pretty displays alignments of several sequences. LineUp is an editor for editing multiple sequence alignments. CompTable helps generate scoring matrices for peptide comparison.

## ALGORITHM

BestFit uses the *local homology* algorithm of Smith and Waterman (Advances in Applied Mathematics 2; 482-489 (1981)) to find the best segment of similarity between two sequences. BestFit reads a scoring matrix that contains values for every possible GCG symbol match (see the LOCAL DATA FILES topic below). The program uses these values to construct a path matrix that represents the entire surface of comparison with a score at every position for the best possible alignment to that point. The *quality* score for the best alignment to any point is equal to the sum of the scoring matrix values of the matches in that alignment, less the gap creation penalty times the number of gaps in that alignment, less the gap extension penalty times the total length of all gaps in that alignment. The gap creation and gap extension penalties are set by you. If the best path to any point has a negative value, a zero is put in that position.

After the path matrix is complete, the highest value on the surface of comparison represents the end of the best region of similarity between the sequences. The best path from this highest value backwards to the point where the values revert to zero is the alignment shown by BestFit. This alignment is the best segment of similarity between the two sequences.

For nucleic acids, the default scoring matrix has a *match* value of 1.0 for each identical symbol comparison and -0.90 for each non-identical comparison (not considering nucleotide ambiguity symbols for this example). The *quality* score for a nucleic acid alignment can, therefore, be determined using the following equation:

$$\begin{aligned} \text{Quality} = & 1.0 \times \text{TotalMatches} + -0.90 \times \text{TotalMismatches} \\ & - (\text{GapCreationPenalty} \times \text{GapNumber}) \\ & - (\text{GapExtensionPenalty} \times \text{TotalLengthOfGaps}) \end{aligned}$$

The *quality* score for a protein alignment is calculated in a similar manner. However, while the default nucleic acid scoring matrix has a single value for all non-identical comparisons, the default protein scoring matrix has different values for the various non-identical amino acid comparisons. The *quality* score for a protein alignment can therefore be determined using the following equation (where  $\text{Total}_{AA}$  is the total number of A-A (Ala-Ala) matches in the alignment,  $\text{CompVal}_{AA}$  is the value for an A-A comparison in the scoring matrix,  $\text{Total}_{AB}$  is the total number of A-B (Ala-Asx) matches in the alignment,  $\text{CompVal}_{AB}$  is the value for an A-B comparison in the scoring matrix, ...):

$$\begin{aligned} \text{Quality} = & \text{CompVal}_{AA} \times \text{Total}_{AA} \\ & + \text{CompVal}_{AS} \times \text{Total}_{AS} \\ & - \text{CompVal}_{AC} \times \text{Total}_{AC} \\ & \vdots \\ & - \text{CompVal}_{XX} \times \text{Total}_{XX} \\ & - (\text{GapCreationPenalty} \times \text{GapNumber}) \\ & - (\text{GapExtensionPenalty} \times \text{TotalLengthOfGaps}) \end{aligned}$$

For a more complete discussion of scoring matrices, see the Data Files manual.

## CONSIDERATIONS

### BestFit Always Finds Something

BestFit always finds an alignment for any two sequences you compare -- even if there is no significant similarity between them! You must evaluate the results critically to decide if the segment shown is not just a random region of relative similarity.

### The Segments Shown Obscure Alternative Segments

BestFit only shows one segment of similarity; so if there are several, all but one is obscured. You can approach this problem with graphic matrix analysis (see the Compare and DotPlot programs). Alternatively, you can run BestFit on ranges outside the ranges of similarity found in earlier runs to bring other segments out of the shadow of the best segment.

### The Best Fit is Only One Member of a Family

Like all fast gapping algorithms, the alignment displayed is a member of the family of best alignments. This family may have other members of equal quality, but will not have any member with a higher quality. The family is usually significantly different for different choices of gap creation and gap extension penalties. See the CONSIDERATIONS topic in the entry for the Gap program in the Program Manual to learn more about how to assign gap creation and gap extension penalties.

### The Surface of Comparison

The magnitude of the computer's job is proportional to the area of the surface of comparison. That area is determined by the product of the lengths of the two sequences compared. BestFit can evaluate a surface of up to 3.5 million elements. This surface would be large enough to compare two sequences approximately 1,870-symbols long, or one sequence 200-symbols long with another sequence 17,500-symbols long. When you have much longer sequences that are known to align well, you can use the command-line option `-LIMIT` to use the surface more efficiently.

### The Public Scoring Matrix for Nucleic Acid Comparisons is Very Stringent

The scoring matrix `swgapdna.cmp` penalizes mismatches -0.9 so the segments found may be very brief. This penalty means that the alignment cannot be extended by three bases to pick one extra match. The scoring matrix used by Smith and Waterman, when local alignments were first described, used -0.333 for the mismatch penalty. You can use Fetch to copy `randomdna.cmp` and rename it `swgapdna.cmp` to use these values, or use `nwsgapdna.cmp`, which has no mismatch penalty at all.

### Rapid Alignment

When possible, BestFit tries to find the optimal alignment very quickly. If this rapid alignment is not unambiguously optimal, BestFit automatically realigns the sequences to calculate the optimal alignment. When this occurs, the monitor of alignment progress on your terminal screen (`Aligning...`) is displayed twice for a single alignment.

## ALIGNING LONG SEQUENCES

This program can align very long sequences if you know roughly where the alignment of interest begins. Run the program with the command line option `-LIMIT`. Then set the starting coordinates for each sequence near the point where the alignment of interest begins and set gap shift limits on each sequence. The program then aligns the sequences from your starting point such that the sequences do not get out of phase by more than the gap shift limits you have set. If you started both sequences at

base number one and set the gap shift limit for sequence one to 100 and for sequence two to 50, then base 350 in sequence one could not be gapped to any base outside of the range from 300 to 450 on sequence two.

If you omit `-LIMIT` on the command line, the program automatically sets gap shift limits if they are needed to allow the alignment of long sequences to proceed. In this case, the program limits the total length of gaps that can be inserted into each sequence and calculates the best alignment within this incomplete, or *limited*, surface of comparison. The program then performs a calculation to determine whether the alignment could possibly be improved if there were no restriction on the total length of gaps in each sequence. If the program cannot rule out this possibility, it displays the message `*** Alignment is not guaranteed to be optimal ***`. Because the criteria used in the calculation for guaranteeing an optimal alignment are very stringent, a limited alignment often may be optimal even if this message is displayed. In any event, the program continues to completion.

## EVALUATING ALIGNMENT SIGNIFICANCE

This program can help you evaluate the significance of the alignment, using a simple statistical method, with the `-RANDOMIZATIONS` command line option. The second sequence is repeatedly shuffled, maintaining its length and composition, and then realigned to the first sequence. The average alignment score, plus or minus the standard deviation, of all randomized alignments is reported in the output file. You can compare this average *quality* score to the quality score of the actual alignment to help evaluate the significance of the alignment. The number of randomizations can be specified along with the `-RANDOMIZATIONS` command line qualifier; the default is 10.

The score of each randomized alignment is reported to the screen. You can use `<Ctrl>C` to interrupt the randomizations and output the results from those randomized alignments that have been completed.

By ignoring the statistical properties of biological sequences, this simple Monte Carlo statistical method may give misleading results. Please see Lipman, D.J., Wilbur, W.J., Smith, T.F., and Waterman, M.S. (Nucl. Acids Res. 12; 215-226 (1984)) for a discussion of the statistical significance of nucleic acid similarities.

## ALIGNMENT METRICS

BestFit and Gap display four figures of merit for alignments: Quality, Ratio, Identity, and Similarity.

The Quality (described above) is the metric maximized in order to align the sequences. Ratio is the quality divided by the number of bases in the shorter segment. Percent Identity is the percent of the symbols that actually match. Percent Similarity is the percent of the symbols that are similar. Symbols that are across from gaps are ignored. A similarity is scored when the scoring matrix value for a pair of symbols is greater than or equal to 0.50, the *similarity threshold*. This threshold is also used by the display procedure to decide when to put a ':' (colon) between two aligned symbols. You can reset it from the command line with the second optional parameter of `-PAIR`. For instance, the expression `-PAIR=1.0, 0.5` would set the similarity threshold to 0.5.

*The similarity and identity metrics are not optimized by alignment programs so they should not be used to compare alignments.*

## PEPTIDE SEQUENCES

If your input sequences are peptide sequences, this program uses a scoring matrix with matches scored as 1.5 and mismatches scored according to the evolutionary distance between the amino acids as measured by Dayhoff and normalized by Gribskov (Gribskov and Burgess Nucl. Acids Res. 14(16); 6745-6763 (1986)).

**RESTRICTIONS**

Input sequences may not be more than 30,000-symbols long. This program cannot evaluate a surface of comparison larger than 5.5 million elements. A 200 x 27,500 comparison is possible, as well as a 2,300 x 2,300 comparison. See the ALIGNING LONG SEQUENCES topic for help in aligning long sequences that would normally exceed the maximum surface of comparison. You can also ask your system manager to increase the maximum surface of comparison if your system has enough virtual memory.

**SEQUENCE TYPE**

The function of BestFit depends on whether your input sequence(s) are protein or nucleotide. Normally the type of a sequence is determined by the presence of either Type: N or Type: P on the last line of the text heading just above the sequence itself. If your sequence(s) are not the correct type, turn to Appendix VI for information on how to change or set the type of a sequence.

**COMMAND-LINE SUMMARY**

All parameters for this program may be put on the command line. Use the option **-CHECK** to see the summary below and to have a chance to add things to the command line before the program executes. In the summary below, the capitalized letters in the qualifier names are the letters that you *must* type in order to use the parameter. Square brackets ([ and ]) enclose qualifiers or parameter values that are optional. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the User's Guide.

Minimal Syntax: % bestfit [-INfile1=]gamma.seq [-INfile2=]alu.seq -Default

**Prompted Parameters:**

-BEGin1=1	-BEGin2=1	beginning of each sequence
-END1=500	-END2=207	end of each sequence
-NOREV1	-NOREV2	strand of each sequence
-GAPweight=5.0		gap creation penalty (3.0 is protein default)
-LENGthweight=0.3		gap extension penalty (0.1 is protein default)
[-OUTfile1=]gamma.pair		output file for alignment

Local Data Files: -DATA=swgapdna.cmp scoring matrix for nucleic acids  
 -DATA=swgappep.cmp scoring matrix for peptides

**Optional Parameters:**

-OUTfile2=gamma.gap	new sequence file for sequence 1 with gaps added
-OUTfile3=alu.gap	" " " " " 2 " " "
-LIMIT1=499 -LIMIT2=206	limit the surface of comparison
-RANDOMizations[=10]	determine average score from 10 randomized alignments
-PAIR=1.0,0.5,0.1	thresholds for displaying ' ', ':', and '.'
-WIDTH=50	the number of sequence symbols per line
-PAGE=60	adds a line with a form feed every 60 lines
-NOBIGGaps	suppresses abbreviation of large gaps with '.'
-HIGHroad	makes the top alignment for your parameters
-LOWroad	makes the bottom alignment for your parameters
-NCSUMmary	suppresses the screen summary

## ACKNOWLEDGEMENTS

Gap and BestFit were originally written for Version 1.0 by Paul Haeberli from a careful reading of the Needleman and Wunsch (J. Mol. Biol. 48; 443-453 (1970)) and the Smith and Waterman (Adv. Appl. Math. 2; 482-489 (1981)) papers.

Limited alignments were designed by Paul Haeberli and added to the Package for Version 3.0. They were united into a single program by Philip Delaquess for Version 4.0. Default gap penalties for protein alignments were modified according to the suggestions of Rechid, Vingron and Argos (CABIOS 5; 107-113 (1989)).

## LOCAL DATA FILES

The files described below supply auxiliary data to this program. The program automatically reads them from a public data directory unless you either 1) have a data file with exactly the same name in your current working directory; or 2) name a file on the command line with an expression like `-DATA1=myfile.dat`. For more information see Chapter 4, Using Data Files in the User's Guide.

If the first sequence you name is a nucleic acid, BestFit uses the scoring matrix in the public file `swgapdna.cmp`. (SW stands for Smith and Waterman.) If the first sequence you name is a peptide sequence, BestFit reads `swgappep.cmp` instead. The presence of these files in your current working directory causes BestFit to read your version instead. (See the Data Files manual for more information about scoring matrices.)

## OPTIONAL PARAMETERS

The parameters and switches listed below can be set from the command line. For more information, see "Using Program Parameters" in Chapter 3, Basic Concepts: Using Programs in the User's Guide.

`-LIMIT1=20` and `-LIMIT2=20`

let you set *gap shift limits* for each sequence. When you already know of a long similarity between two sequences you can "zip" them together using this mode. The beginning coordinates for each sequence must be near the beginning of the alignment you want to see. The alignment continues so that gaps inserted do not require the sequences to get out of step by more than the gap shift limits. You can align very long sequences rapidly. The surface of comparison is still limited to 3.5 million. The size of a comparison can be predicted by multiplying the average length of the two sequences by the sum of the two shift limits.

If you add `-LIMIT` to the command line without any qualifier value, the program prompts you to enter gap shift limits for each sequence.

`-RANDOMIZATIONS=10`

reports the average alignment score and standard deviation from 10 randomized alignments in which the second sequence is repeatedly shuffled, maintaining the length and composition of the original sequence, and then aligned to the first sequence. You can use the optional parameter to set the number of randomized alignment to some number other than 10.

`-OUTFILE2=seqname1.gap` `-OUTFILE3=seqname2.gap`

This program can write three different output files. The first displays the alignment of sequence one with sequence two. The second is a new sequence file for sequence one, possibly expanded by gaps to make it align with sequence two. The third, like the second, is a new sequence file for sequence two, possibly expanded by gaps to make it align with sequence one. The program writes only the first file unless there are output file options on the command line. If there are any output files named on the command line, only those output files are written. If you add

Aligned sequences (in sequence files) can be displayed with GapShow. Their similarity can be displayed with PlotSimilarity.

The paired output file from this program displays sequence similarity by printing one of three characters between similar sequence symbols: a pipe character (|), a colon (:), or a period (.). Normally a pipe character is put between symbols that are the same, a colon is put between symbols whose comparison value is greater than or equal to 0.50, and a period is put between symbols whose comparison value is greater than or equal to 0.10. You can change these *match display thresholds* from the command line. The three parameters for `-PAIR` are the display thresholds for the pipe character, colon, and period. The match display criterion for a pipe character changes from symbolic identity (the default) to the quantitative threshold you have set in the first parameter. A pipe character will no longer be inserted between identical symbols unless their comparison values are greater than or equal to this threshold. If you still want a pipe character to connect identical symbols, use `x` instead of a number as the first parameter. (See the *Data Files* manual for more information about scoring matrices.)

When you print the output from this program, it may cross from one page to another in a frustrating way – especially when you print on individual sheets. This option adds form feeds to the output file in order to try to keep clusters of related information together. You can set the number of lines per page by supplying a number after the `-PAGE` qualifier.

puts 50 sequence symbols on each line of the output file. You can set the width to anything from 10 to 150 symbols.

suppresses large gap abbreviations, showing all the sequence characters across from large gaps. Usually, gaps that extend one sequence by more than one complete line of output are abbreviated with three dots arranged in a vertical line.

The insertion of gaps is, in many cases, arbitrary, and equally optimal alignments can be generated by inserting gaps differently. When equally optimal alignments are possible, this program can insert the gaps differently if you select either the `-LOWroad` or the `-HIGHroad` options. Here are examples for the alignment of GACCAT with GACAT with different parameters.

```
LowRead: 1 GACCAT 6
          |||
          1 GA.CAT 5      Quality = 4.0

HighRead: 1 GACCAT 6
           ||| ||
           1 GAC.AT 5     Quality = 4.0
```



For: Match = 1.0 MisMatch = 0.0  
Gap weight = 3.0 Length Weight = 0.0

HighRoad: 1 GACCAT 6  
          111 Quality = 3.0  
          1 GACAT. 5

LowRoad: 1 GACCAT 6  
          111 Quality = 3.0  
          1 .GACAT 5

Essentially the *low road* shifts all of the arbitrary gaps in sequence two to the left and all of the arbitrary gaps in sequence one to the right. The *high road* does exactly the opposite. When neither *high road* nor *low road* is selected, the program tries not to insert a gap whenever that is possible and uses the *high road* alternative for all collisions.

#### -SUMmary

writes a summary of the program's work to the screen when you've used the -Default qualifier to suppress all program interaction. A summary typically displays at the end of a program run interactively. You can suppress the summary for a program run interactively with -NOSUMmary.

Use this qualifier also to include a summary of the program's work in the log file for a program run in batch.

Printed: July 13, 1995 08:19 (1162)